1. Consider the dataset related to occurence of oral leukoplakia leukoplakia.txt:

```
Leukoplakia Alcohol Smoker
          0       1      0
          0       1      0
          0       1      0
          0       1      0
          0       1      0
          0       1      0
            .       .      .
          1       4      1
```

```
The leukoplakia data is about occurence of oral leukoplakia with
explanatory variables being smoking and alcohol consumption.

Leukoplakia
Has the person oral leukoplakia? yes = 1, no = 0
Alcohol
How much alcohol did the person drink on average? no = 1, less than 40g = 2,
less than 80g = 3, more than 80g = 4
Smoker
Smoker? yes = 1, no = 0
```

Denote the variables as following

$$Y = \text{Leukoplakia}, \quad X_1 = \text{Alcohol}, \quad X_2 = \text{Smoker}.$$

(a) Assume $Y_i \sim Ber(\mu_{jh})$. Consider the model

$$\text{logit}(\mu_{jh}) = \beta_0 + \beta_j + \alpha_h,$$

where indexes $j, h$ are related in order to categories of variables $X_1, X_2$. Calculate the maximum likelihood estimate for the probability $P(Y_{i_*} = 1)$ when

$$x_{i_*1} = 3 = \text{"less than 80g"}, \qquad x_{i_*2} = 1 = \text{"yes"}.$$

(2 points)

(b) Consider the model

$$\text{logit}(\mu_{jh}) = \beta_0 + \beta_j + \alpha_h.$$

Test at 5% significance level, is the explanatory variable $X_2$ =Smoker statistically significant variable. Calculate the value of the test statistic.

(1 point)

(c) Consider the model

$$\text{logit}(\mu_{jh}) = \beta_0 + \beta_j + \alpha_h.$$

Calculate the estimate for the odd ratio

$$\psi_{\text{more than 80g,yes|less than 80g,yes}}.$$

(1 point)

(d) Assume $\text{Var}(Y_i) = \phi\mu_{jh}(1 - \mu_{jh})$. Consider the hypotheses

$$H_0 : \text{Model } \text{logit}(\mu_{jh}) = \beta_0 + \beta_j + \alpha_h \text{ is the true model},$$
$$H_1 : \text{Model } \text{logit}(\mu_{jh}) = \beta_0 + \beta_j + \alpha_h + \gamma_{jh} \text{ is the true model}.$$

Select the appropriate test statistic to test the above hypotheses. Calculate the value of the test statistic.

(1 points)

(e) Which link function fits best to data in your opinion, if you use the main effect model

$$g(\mu_{jh}) = \beta_0 + \beta_j + \alpha_h$$

to model the data?

i. Probit link $g(\mu_i) = \Phi^{-1}(\mu_i)$,
ii. Cauchy link $g(\mu_i) = F^{-1}_{\text{cauchy}}(\mu_i)$,
iii. Gumbel link (complementary log-log) $g(\mu_i) = \log\left(-\log(1 - \mu_i)\right)$,

(1 point)

2. Consider the data set appleCRA7152.txt, where it has been studied how the probability of bacterial spores of Alicyclobacillus Acidoterrestris CRA7152 growing in apple juice depends on the properties of the apple juice.

```
    pH Nisin Temperature Brix Growth
1  5.5    70          50   11      0
2  5.5    70          43   19      0
3  5.5    50          43   13      1
4  5.5    50          35   15      1
5  5.5    30          35   13      1
.
73 5.5    70          50   19      0
74 3.5     0          25   11      0

Presence/Absence of growth of CRA7152 in apple juice
as a function of pH (3.5-5.5), Brix (11-19), temperature (25-50C),
and Nisin concentration (0-70)

X1=pH
X2=Nisin concentration
X3=Temperature
X4=Brix Concentration
Y=Growth    (1=Yes, 0=No)

Source: W.E.L. Pena, P.R. De Massaguer, A.D.G. Zuniga, and S.H. Saraiva (2011).
"Modeling the Growth Limit of Alicyclobacillus Acidoterrestris CRA7152
in Apple Juice: Effect of pH, Brix, Temperature, and Nisin Concentration,"
Journal of Food Processing and Preservation, Vol. 35, pp. 509-517.
```

Denote the variables as following:

$$Y = \text{Growth}, \quad X_1 = \text{pH}, \quad X_2 = \text{Nisin}, \quad X_3 = \text{Temperature}, \quad X_4 = \text{Brix}.$$

(a) Consider modeling the expected value of the response variable $Y = \mathsf{Growth}$ by the explanatory variables $X_1, X_2, X_3, X_4$. Select the appropriate default distribution for the response variable $Y$, and consider several competing models. Choose the model which you feel is the most suitable one for modeling the expected value of the response variable $Y = \mathsf{Growth}$. Not all explanatory variables $X_1, X_2, X_3, X_4$ need to be included into your final model. Which link function $g(\mu_i)$ you chose for you model?

i. Identity link $g(\mu_i) = \mu_i$,
ii. $\log$ link $g(\mu_i) = \log(\mu_i)$,
iii. Inverse link $g(\mu_i) = \frac{1}{\mu_i}$,
iv. $\mathrm{logit}$ link $g(\mu_i) = \mathrm{logit}(\mu_i)$,
v. Probit link $g(\mu_i) = \Phi^{-1}(\mu_i)$,
vi. Cauchy link $g(\mu_i) = F^{-1}_{\mathsf{cauchy}}(\mu_i)$,
vii. Gumbel link $g(\mu_i) = \log\left(-\log(1-\mu_i)\right)$,

(2 points)

(b) Based on your chosen model, calculate the maximum likelihood estimate for the expected value $\mu_i$ when the explanatory variables are set on values

$$X_1 = 4.5, \quad X_2 = 20, \quad X_3 = 30, \quad X_4 = 17.$$

(1 point)

(c) Based on your chosen model, calculate the 95% confidence interval estimate for the expected value $\mu_i$ when the explanatory variables are set on values

$$X_1 = 4.5, \quad X_2 = 20, \quad X_3 = 30, \quad X_4 = 17.$$

(1 point)

(d) Let us assume that there are 100 apple juices with explanatory variables are set on values

$$X_1 = 4.5, \quad X_2 = 20, \quad X_3 = 30, \quad X_4 = 17.$$

How many of these 100 juices are such that bacterial spores of Alicyclobacillus Acidoterrestris CRA7152 are occuring in them? Create 80% prediction interval for the number of apple juices affected by Alicyclobacillus Acidoterrestris CRA7152 bacteria.

(2 points)

3. (a) Let $Y_i \sim Cat(\theta_{i1}, \theta_{i2}, \theta_{i3})$, and consider the multinomial logit models

$$\log\left(\frac{\theta_{i2}}{\theta_{i1}}\right) = \mathbf{x}_i'\boldsymbol{\beta}_2,$$
$$\log\left(\frac{\theta_{i3}}{\theta_{i1}}\right) = \mathbf{x}_i'\boldsymbol{\beta}_3,$$

where $\theta_{i1} + \theta_{i2} + \theta_{i3} = 1$. Show that

$$\theta_{i1} = \frac{1}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}_2} + e^{\mathbf{x}_i'\boldsymbol{\beta}_3}},$$
$$\theta_{i2} = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}_2}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}_2} + e^{\mathbf{x}_i'\boldsymbol{\beta}_3}},$$
$$\theta_{i3} = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}_3}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}_2} + e^{\mathbf{x}_i'\boldsymbol{\beta}_3}}.$$

(2 points)

(b) Let the random variable $Y_i$ be defined on ordinal scale with $m$ distinctive possible outcomes. Let the possible outcomes have natural order "1" $<$ "2" $<$ "3". Consider cumulative proportional odds logit model

$$\log\left(\frac{P(Y_i \leq k)}{1 - P(Y_i \leq k)}\right) = \text{logit}\left(P(Y_i \leq k)\right) = \beta_{0k} + \beta_1 x_{i1}, \qquad k = 1, 2.$$

Solve the probabilities $P(Y_i = 1), P(Y_i = 2), P(Y_i = 3)$ as functions of parameters $\beta_{0k}, \beta_1$.

(2 points)

(c) In generalized linear models, the likelihood equations can written in form

$$\frac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\text{Var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) = 0, \qquad j = 0, 1, 2 \ldots p.$$

Consider now the simple logit model with

$$Y_i \sim Ber(\mu_i),$$
$$\text{logit}(\mu_i) = \eta_i = \beta_0.$$

What kind of more simplified form the likelihood equations have in this case? That is, what form $\frac{\partial l(\beta_0)}{\partial \beta_0}$ has in the simple logit model? By using the likelihood equations, find the maximum likelihood estimator $\hat{\beta}_0$.

(2 points)